

## Supplementary Tables and Figures

<b>Table S1.</b> Average number of reads in the top 20 most abundant genera for TCGA-BLCA as computed in the Poore <i>et al.</i> study, averaged over 129 WGS primary tumor and 27 solid tissue normal samples.		
<b>Genus</b>	<b>Average read count per sample, Poore <i>et al.</i></b>	<b>Average read count per sample, this study</b>
<i>Streptococcus</i>	560212	36
<i>Mycobacterium</i>	410968	6.1
<i>Staphylococcus</i>	240880	282
<i>Waddlia</i>	55280	0
<i>Bacillus</i>	53659	4
<i>Escherichia</i>	50207	3.4
<i>Bordetella</i>	46075	0.4
<i>Pseudomonas</i>	44186	133
<i>Pseudoalteromonas</i>	40837	0.2
<i>Vibrio</i>	34216	0.2
<i>Streptomyces</i>	28317	2.9
<i>Piscirickettsia</i>	24818	0
<i>Klebsiella</i>	21069	2.1
<i>Microbacterium</i>	16773	0.9
<i>Xanthomonas</i>	15681	22
<i>Shigella</i>	14208	0.7
<i>Acinetobacter</i>	12562	8.6
<i>Bacteroides</i>	12013	7.7
<i>Neisseria</i>	11988	0.4
<i>Salmonella</i>	11234	0.3

<b>Table S2.</b> Average number of reads in the top 20 most abundant genera for TCGA-BLCA as measured in this study, averaged over 129 WGS primary tumor and 27 solid tissue normal samples.		
<b>Genus</b>	<b>Average read count per sample, this study</b>	<b>Average read count per sample, Poore <i>et al.</i></b>
<i>Enterococcus</i>	447	9899
<i>Veillonella</i>	385	726
<i>Staphylococcus</i>	282	240880
<i>Aerococcus</i>	165	1756
<i>Pseudomonas</i>	133	44186
<i>Peptoniphilus</i>	95	737
<i>Finegoldia</i>	80	258
<i>Stenotrophomonas</i>	64	846
<i>Anaerococcus</i>	50	2451
<i>Prevotella</i>	49	1062
<i>Streptococcus</i>	36	560212
<i>Cupriavidus</i>	36	137

<i>Actinotignum</i>	31	153
<i>Citrobacter</i>	25	695
<i>Cutibacterium</i>	24	0
<i>Betapolyomavirus</i>	22	0
<i>Xanthomonas</i>	22	15681
<i>Erysipelatoclostridium</i>	22	1.1
<i>Campylobacter</i>	18	2185
<i>Eimeria</i>	12	0

**Table S3.** Average read counts for the top 20 genera, ranked by weights assigned by the machine learning classifier using the "all putative contaminants removed" dataset and classifying BLCA primary tumor samples versus all other tumor types. Counts are averaged over 129 WGS primary tumor and 27 solid tissue normal samples.

<b>Genus</b>	<b>Average read count, Poore <i>et al.</i></b>	<b>Average read count, this study</b>
<i>Nitrospira</i>	7.5	0
<i>Elizabethkingia</i>	460	0.1
<i>Leptospira</i>	3053	0
<i>Campylobacter</i>	2185	18
<i>Histophilus</i>	162	0
<i>Capnocytophaga</i>	56	0.2
<i>Chelativorans</i>	0	0
<i>Sediminibacterium</i>	2.6	0
<i>Scardovia</i>	0.6	0
<i>Lysobacter</i>	47	0.2
<i>Stomatobaculum</i>	0	0
<i>Gallibacterium</i>	20	0
<i>Turicella</i>	0.1	0
Betaretrovirus	0.2	0
<i>Exiguobacterium</i>	789	0
<i>Wolbachia</i>	291	0
<i>Bacteroides</i>	12013	7.7
Alphapapillomavirus	403	0.1
<i>Hydrogenibacillus</i>	17	0
<i>Candidatus Stoquefichus</i>	75	0

**Table S4.** Average read counts for the top 20 genera, ranked by weights assigned by the machine learning classifier using the "all putative contaminants removed" dataset and classifying BLCA primary tumor samples versus solid tissue normal samples. Counts are averaged over 129 WGS primary tumor and 27 solid tissue normal samples.

<b>Genus</b>	<b>Average read count, Poore <i>et al.</i></b>	<b>Average read count, this study</b>
--------------	--	---------------------------------------

<i>Lachnoclostridium</i>	18	0.2
<i>Methylobacter</i>	1.7	0
<i>Marinitoga</i>	157	0
<i>Vibrio</i>	34216	0.2
<i>Crocinitomix</i>	1.3	0
<i>Flammeovirga</i>	2.3	0
<i>Desulfobacter</i>	0.4	0
<i>Paeniclostridium</i>	148	0.1
<i>Aliihoeflea</i>	0	0
<i>Cellulomonas</i>	73	0.1
<i>Candidatus Nitrosopelagicus</i>	0.2	0
<i>Tymovirus</i>	10	0
<i>Betapartitivirus</i>	72	0
<i>Microvirga</i>	77	0
<i>Salsuginibacillus</i>	0.6	0
<i>Nepovirus</i>	5.7	0
<i>Aeromonas</i>	435	1.9
<i>Klebsiella</i>	21069	2.1
<i>Pusillimonas</i>	0.5	0
<i>Candidatus Evansia</i>	1.7	0

**Table S5.** Average number of reads in the top 20 most abundant genera for TCGA-HNSC as computed in the Poore *et al.* study, averaged over 334 WGS samples.

<b>Genus</b>	<b>Average read count, Poore <i>et al.</i></b>	<b>Average read count, this study</b>
<i>Streptococcus</i>	1335308	2041
<i>Mycobacterium</i>	1001984	22.9
<i>Staphylococcus</i>	670494	121
<i>Pseudomonas</i>	273551	5685
<i>Escherichia</i>	212531	64.4
<i>Mesorhizobium</i>	158289	115
<i>Waddlia</i>	144680	0.0
<i>Bacillus</i>	137082	11.6
<i>Neisseria</i>	128526	1651
<i>Pseudoalteromonas</i>	116176	1.0
<i>Streptomyces</i>	96294	16.8
<i>Vibrio</i>	93027	2.8
<i>Bordetella</i>	84532	5.3
<i>Shigella</i>	70859	1.6
<i>Piscirickettsia</i>	66821	0.0
<i>Treponema</i>	63757	4774
<i>Fusobacterium</i>	57873	10003

<i>Salmonella</i>	53463	1.2
<i>Klebsiella</i>	51586	34.5
<i>Microbacterium</i>	50001	10.1

**Table S6.** Average number of reads in the top 20 most abundant genera for TCGA-HNSC as measured in this study, averaged over 334 WGS samples.

<b>Genus</b>	<b>Average read count, this study</b>	<b>Average read count, Poore <i>et al.</i></b>
<i>Fusobacterium</i>	10003	57873
<i>Capnocytophaga</i>	5981	13925
<i>Prevotella</i>	5706	47835
<i>Pseudomonas</i>	5685	273551
<i>Treponema</i>	4774	63757
<i>Campylobacter</i>	2102	26697
<i>Streptococcus</i>	2041	1335308
<i>Neisseria</i>	1651	128526
<i>Veillonella</i>	1328	3826
<i>Haemophilus</i>	1270	6870
<i>Sphingomonas</i>	1094	12979
<i>Leptotrichia</i>	721	1893
<i>Stenotrophomonas</i>	590	2646
<i>Parvimonas</i>	546	4062
<i>Tannerella</i>	428	2887
<i>Porphyromonas</i>	410	17431
<i>Selenomonas</i>	349	1111
<i>Rothia</i>	331	1027
<i>Bifidobacterium</i>	323	1522
<i>Bradyrhizobium</i>	301	2852

**Table S7.** Average read counts for the top 20 genera, ranked by weights assigned by the machine learning classifier using the "all putative contaminants removed" dataset and classifying HNSC primary tumor samples versus all other tumor types. Counts are averaged over 170 WGS primary tumor, 140 blood derived normal, and 24 solid tissue normal samples.

<b>Genus</b>	<b>Average read count, Poore <i>et al.</i></b>	<b>Average read count, this study</b>
--------------	--	---------------------------------------

<i>Microvirga</i>	209	1.8
<i>Gemmata</i>	13385	0.4
<i>Desulfomicrobium</i>	1.7	0.8
<i>Exiguobacterium</i>	1331	0.3
<i>Nitrosopelagicus</i>	0.4	0.0
Alphapapillomavirus	8577	0.4
<i>Phenylobacterium</i>	21	4.0
<i>Klebsiella</i>	51586	35
<i>Plantibacter</i>	0.3	0.3
<i>Terracoccus</i>	5.0	0.0
<i>Marichromatium</i>	0.8	0.3
Betapartitivirus	159	0.0
<i>Gottschalkia</i>	2.2	0.4
<i>Acidithiobacillus</i>	142	0.4
<i>Desulfococcus</i>	24868	0.2
<i>Epilithonimonas</i>	5.8	0.0
<i>Luteibacter</i>	1036	0.5
<i>Spiroplasma</i>	356	0.7
<i>Mannheimia</i>	429	2.2
<i>Roseivirga</i>	4.2	0.1

**Tables S8, S9, and S10** (separate files): These tables contains all read counts, reported at the genus level, from the bladder cancer (BLCA), head and neck cancer (HNSC), and breast cancer (BRCA) samples from TCGA. The tables include read counts for bacteria, archaea, and viruses. All samples were classified against a KrakenUniq database as described in Methods. **Table S8** (156 x 1063) has read counts for 156 BLCA samples, including 129 primary tumor and 27 solid tissue normal samples, filtered to remove human reads as described in Methods.

Overall, 1,063 genera were identified (i.e., contained at least one non-zero count) in the 156 samples. **Table S9** (334 x 1573) contains read counts for 334 HNSC WGS, including 170 primary tumor samples, 140 blood derived normal samples, and 24 solid tissue normal samples, filtered to remove human reads as described in Methods. Overall, 1,573 genera were identified (i.e., contained at least one non-zero count) across the 334 samples. **Table S10** (238 x 1,200) contains read counts for 238 BRCA WGS samples, including 114 primary tumor samples, 106 blood derived normal samples, 16 solid tissue normal samples, and 2 metastatic samples, filtered to removed human reads as described in Methods. Overall, 1,200 genera were identified (i.e., contained at least one non-zero count) across the 238 samples.

**Supplementary Data File 1.** List of all species contained in the Kraken database used in this study.

**Supplementary Data File 2.** List of all genera contained in the Kraken database used in this study.



